Biometrics 69, 1–5
March 2013

DOI: 10.1111/j.1541-0420.2005.00454.x

Parameter Estimation in Cox Proportional Hazard Models with Missing Censoring Indicators

Naomi Brownstein[1,*], Jianwen Cai[1,], Gary Slade[2,***], Eric Bair[1,2,****]**

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill NC, U.S.A.

[2]School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill NC, U.S.A.

**email:* nbrownst@mail.bios.unc.edu

***email:* cai@bios.unc.edu

****email:* gary_slade@dentistry.unc.edu

*****email:* ebair@email.unc.edu

Summary: In a prospective cohort study, examining all participants for incidence of the condition of interest may be prohibitively expensive. For example, the “gold standard” for diagnosing temporomandibular disorder (TMD) is a clinical examination by an expert dentist. In a large study, examining all subjects in this manner is infeasible. Instead, it is common to use a cheaper (and less reliable) examination to screen for possible incident cases and perform the “gold standard” examination only on participants who screen positive on this simpler examination. Unfortunately, some subjects may leave the study before receiving the “gold standard” examination. Within the framework of survival analysis, this results in missing censoring indicators. Motivated by the Orofacial Pain: Prospective Evaluation and Risk Assessment(OPPERA) study, a large cohort study of TMD, we propose a method for parameter estimation in survival models with missing censoring indicators. We estimate the probability of being a case for those with no “gold standard” examination using logistic regression. These predicted probabilities are used to generate multiple imputations of each missing case status and estimate the hazard ratios associated with each putative risk factor. The variance introduced by the procedure is estimated using multiple imputation. We simulate data with missing censoring indicators and show that our method performs as well as or better than the competing methods. Finally, we apply the proposed method to analyze data from the OPPERA study.

Key words: Cox regression; missing data; multiple imputation; survival analysis.

1. Introduction

Time-to-event analyses are frequently conducted in medicine, actuarial science, and numerous other fields of applied science. There is a well-developed set of survival analysis methods implemented in standard software. Semi-parametric methods, such as the Cox proportional hazards model, allow robust estimation of the effects of covariates on the hazard function. Yet, these methods require the analyst to know not only the time of the event or censoring, but also the event status, which may not always be available.

Often, the outcome of interest may be difficult to ascertain. For example, in oncology studies, researchers may want to differentiate between deaths due to cancer and deaths due to car accidents or other unrelated causes. Investigators may easily record the mortality of all subjects, but it may be extremely difficult or costly to find out exactly why each subject died. One possible solution to this problem is delayed event adjudication (Cook and Kosorok, 2004). This means that possible cases are not identified immediately but screened using simple methods that may have poor sensitivity or specificity; then, one or more experts examines the screened candidate cases using a more precise, but also more costly and time-consuming method to determine the true event status.

The study that motivates our work is Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA), a prospective cohort study to identify risk factors for the onset of temporomandibular disorders (TMD). A total of 3,263 initially TMD-free subjects were recruited at four study sites between 2006 and 2008. A specialized dental examination based on the Research Diagnostic Criteria (RDC) (Dworkin and LeResche, 1992) is required for accurately diagnosing TMD. It is impractical to perform an RDC exam on every participant in such a large study, especially if most study participants are unlikely to have the condition. Instead, this “gold standard” examination is performed only on subjects who screen positive on a quarterly questionnaire designed to assess recent orofacial pain (Bair et al., 2013a).

However, some subjects who screen positive on the questionnaire may be lost to follow-up before receiving the “gold standard” examination. A time-to-event analysis would then have some subjects with missing censoring indicators.

Cook and Kosorok (2004) estimate parameters in Cox proportional hazard models with missing censoring indicators by weighting observations by their probability of being a true case and show that the estimators are consistent and asymptotically normally distributed. However, the standard error of their proposed estimate cannot be easily obtained using existing software without bootstrapping. For the OPPERA data, a separate Cox model will be calculate for each putative risk factor of interest, including approximately 3,000 genetic markers. Consequently, applying this method to the OPPERA genetic data would be computationally intractable. Moreover, the method of Cook and Kosorok (2004) does not apply to incidence rate estimation, which is of interest in the OPPERA study.

The likelihood that a participant will come in for their RDC exam differs based on known factors such as gender, race, or socio-economic status (Bair et al., 2013b). This indicates that the censoring indicators in the OPPERA study may not be missing completely at random (MCAR). Application of models that assume MCAR censoring indicators may result in biased estimates of hazard ratios for covariates of interest. More significantly, a subject’s responses to their QHU are predictive of whether or not they are a case. This setting presents statistical challenges, which require care in order to avoid bias and maintain efficiency. There is a clear need for new methodology to effectively answer the research questions of the OPPERA study.

In this paper, we propose a method for parameter and variance estimation in the Cox regression model with missing censoring indicators. We describe our method in section 2. In section 3, we report the results of simulations. Finally, in section 4 we apply our method to the OPPERA study. We conclude with a discussion in section 5.

2. Model

Assume there are n independent subjects. For each subject, $i = 1, \ldots, n$, let C_i and T_i denote the potential times until censoring and failure, respectively, $V_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leqslant C_i)$, $N_i(t) = I(T_i < t)$, Z_i a $p \times 1$ vector of covariates and X_i a $q \times 1$ vector of auxiliary covariates. We assume the hazard for subject i follows a Cox proportional hazards model

$$\lambda(t|z_i) = \lambda_0(t) \exp(\beta' z_i) \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function. Let ξ_i denote the indicator that Δ_i is observed and $\sigma_i = \xi_i \Delta_i$. We observe (V_i, ξ_i, σ_i) for $i = 1, \ldots, n$.

In the OPPERA study, V_i is the length of time since subject i is enrolled in the study until either he or she screens positive on a QHU and is subsequently diagnosed with TMD or is lost to follow-up. (Note that the study protocol requests that subjects fill out QHUs until the end of the study, even if the subject is diagnosed with TMD. For the purposes of this paper, we do not consider QHUs for subjects who have already been diagnosed with TMD.) If subject i screened positive on a QHU and subsequently had a positive RDC exam, then $\Delta_i = 1$. If subject i either screened negative on the last QHU before loss-to-follow up or screened positive on the QHU and had a negative RDC exam, then $\Delta_i = 0$. If subject i screened positive on the last QHU but did not come in for the RDC exam, then Δ_i is missing and $\xi_i = 0$. If Δ_i is observed, then $\xi_i = 1$. Measured when a participant enrolled in the study, putative risk factors for TMD are denoted by Z_i, and a participant's responses to their QHU are denoted by X_i. For OPPERA, we also define $Q_i = 1$ if subject i screens positive on the last QHU before either a positive diagnosis of TMD or loss to follow-up and $Q_i = 0$ otherwise.

We assume the censoring indicators are missing at random (MAR) as follows:

$$\begin{aligned} & P(\xi_i = 0|X_i, V_i, \Delta_i, Q_i = 1) \qquad (2) \\ = \; & P(\xi_i = 0|X_i, V_i, Q_i = 1) = \rho(X_i, V_i) \end{aligned}$$

We model the probability that subject i with a missing censoring indicator is a case by a logistic regression model based on X_i and V_i, i.e.

$$\begin{aligned} & p(\Delta_i = 1 | X_i, V_i, Z_i, \xi_i = 0, Q_i = 1) \\ = \; & p(\Delta_i = 1 | X_i, V_i, Q_i = 1) \\ = \; & \frac{\exp(\alpha' X_i + \gamma V_i)}{1 + \exp(\alpha' X_i + \gamma V_i)} I(Q_i = 1) \end{aligned} \qquad (3)$$

That is, we estimate the probability that a subject who did not come in for the RDC exam actually has TMD based on the time of the QHU and their answers on the QHU. In the OPPERA study, we observed empirically that the probability of having developed TMD was not associated with baseline covariates once we controlled for the responses on the QHU.

Then, for those individuals who screened positive on the last QHU (i.e. those with $Q_i = 1$) who did not come in for their RDC exam, the estimated probability of being a case is estimated by (3) with the parameters replaced by their respective estimates based on individuals who were examined.

Note, if there are repeated measures, we may use a generalized linear mixed model rather than a logistic regression model. For example, if a majority of subjects repeatedly screen positive on their QHUs and come in for at least one RDC exam, then we would have multiple observations per subject. In that case, fitting a mixed model rather than a logistic regression model would account for correlations between responses of the same subject.

Once we have the predicted failure probabilities, we may use them to create imputed datasets. We then use the completed data to fit a standard Cox proportional hazards model and estimate the parameters and their standard errors via multiple imputation.

2.1 *Multiple Imputation*

One popular method of parameter and standard error estimation is multiple imputation. Widely available in statistical software, multiple imputation is commonly used by both statisticians and non-statisticians alike. For a comprehensive reference on multiple imputation, see the review paper of Rubin (1996). Our imputation procedure is as follows:

(i) Estimate predicted probabilities as described in the previous section.

(ii) Retain each observation with observed censoring indicators.

(iii) For each observation with a missing censoring indicator, generate a Bernoulli random variable with success probability equal to the predicted probability found in step (i).

(iv) Combine the raw and imputed data from steps (ii) and (iii) to form a completed dataset.

(v) Fit the Cox proportional hazards model to the completed dataset.

(vi) Record each parameter estimate $\hat{\beta}_j$ and covariance matrix $\hat{U}_j$.

(vii) Repeat steps (ii)-(vi) m times.

Next, we combine all of the estimates. The average parameter estimate is

$$\bar{\beta} = \frac{1}{m} \sum_{j=1}^{m} \hat{\beta}_j, \tag{4}$$

the within-imputation variance estimate is

$$\bar{U} = \frac{1}{m} \sum_{j=1}^{m} \hat{U}_j, \tag{5}$$

and the between-imputation variance

$$\hat{B} = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{\beta}_j - \bar{\beta})(\hat{\beta}_j - \bar{\beta})'. \tag{6}$$

Finally, the estimated covariance matrix is

$$V\hat{a}r(\bar{\beta}) = \bar{U} + (1 + \frac{1}{m})\hat{B}. \tag{7}$$

2.2 *Estimation of Incidence*

We can also estimate incidence rates using Poisson regression instead of Cox regression. In order to estimate incidence, we estimate the predicted case probabilities as described previously. Then we impute case status as described in section 2.1 but we fit Poisson regression models, rather than Cox models, to the completed datasets. Finally, we calculate the incidence rate based on the estimates of the regression coefficients in the Poisson model.

3. Simulations

We simulated data with missing censoring indicators and compared methods with respect to bias, coverage, and confidence interval width. Survivaltimes for 1,000 individuals were generated with exponentially distributed failure times under a proportional hazards model with covariates as proposed by Bender et al. (2005). That is, the survival time for each individual was distributed according to equation (1) where $\lambda_0(t) = 1$ is the baseline hazard. For our simulations, the $Z_i = X_{i1}$ was a single covariate following a normal distribution with mean 2 and unit variance. In other words, the failure times, T_i followed an exponential distribution with hazard $\exp(\beta' X_{i1})$. We used $\beta \in \{-0.5, -1.5, -3\}$. The censoring times, C_i followed an exponential distribution with mean 5. This yielded about 35%, 75% and 90% censoring, respectively.

For each observation, we generated a random normal variable, $X_{i2} \sim N(\Delta_i, 0.3)$. Larger values of X_{i2} were more likely to be associated with a failure event. We use X_{i2}, a continuous measure of the potential of being as a case, to generate the indicator $Q_i = I(X_{i2} > 0.5)$.

We created missing censoring indicators under the following classical missing data mechanisms of Rubin (1976):

(I) The probability of having a missing censoring indicator is independent of the data. This is known as missing completely at random (MCAR).

(II) The probability of having a missing censoring indicator depends on an observed covariate. This is known as missing at random (MAR).

(III) The probability of having a missing censoring indicator depends on the censoring indicator. This is known as missing not at random (MNAR).

Our method assumes data are missing according to mechanism II. In order to parallel the setup of OPPERA, for mechanisms II and III, observations were potentially missing if $Q_i = 1$. Details and results for missing data mechanisms I and III are shown in Web Appendix A. For mechanism II, we set censoring indicators to be missing based on x_{i1} and V_i, i.e. they were set to missing with probability

$$\begin{aligned} \rho_i(X_i, V_i) &= P(\xi_i = 0|X_i, V_i, Q_i = 1) \\ &= \frac{exp(-0.2 - 0.3x_{i1} + 0.1V_i)}{1 + exp(-0.2 - 0.3x_{i1} + 0.1V_i)}. \end{aligned} \tag{8}$$

Within the framework of OPPERA, for person i, $V_i = min(T_i, C_i)$ corresponds to the time of the last positive QHU before a positive diagnosis of TMD or loss to follow-up, $\xi_i = I(\Delta_i \text{ is observed})$ to the indicator of whether person i came in for their RDC exam if $Q_i = 1$, Δ_i to the indicator of whether person i was diagnosed with TMD, X_{i1} to a risk factor for TMD measured at baseline, X_{i2} to a covariate collected on the last QHU, and Q_i as an indicator of whether the person screened positive on their last QHU.

In each simulated dataset, we first fit all observations with observed censoring indicators who had screened positive on their QHUs to a logistic regression model for case status with the covariates X_{i1} and X_{i2}. That is, using the complete data (i.e. observations with $Q_i = 1$ and $\xi_i = 1$), we fit the logistic regression model for the event probability conditional on $X_i' = (1, X_{i1}, X_{i2})$ and V_i, namely

$$logit\{Pr(\Delta_i = 1|X_i, V_i, Q_i = 1, \xi_i = 1)\} = \alpha' X_i + \gamma V_i \tag{9}$$

We output the predicted probabilities, $\hat{p}_i = \frac{\hat{\alpha}' X_i + \hat{\gamma} V_i}{1 + \hat{\alpha}' X_i + \hat{\gamma} V_i}$, for individuals with $Q_i = 1$.

To evaluate the performance of our method, we used multiple imputation to create 100 datasets for each simulation as described in Section 2.1. That is, we generated 100 datasets with identical data for X_i, T_i, C_i, ξ_i. For each observation i with $Q_i = 1$ and $\xi_i = 0$, we generated failure indicators $\hat{\Delta}_{ij} \sim Bernoulli(\hat{p}_i)$ independently for each imputation j.

We fit a Cox proportional hazards model to each dataset completed by multiple imputation and recorded the multiple imputation estimates of the regression coefficient and its variance. These were aggregated using equations (4) and (7) to create confidence intervals for the multiple imputation estimates.

We compared the performance of our method with the performance of the method of Cook and Kosorok (2004). To obtain the estimates of Cook and Kosorok (2004), for each simulated data set, we calculated the estimated probabilities $\hat{p}_i$ that (unobserved) event for subject i is an event, as described previously. We then fit a weighted Cox proportional hazards model to the dataset. For observations with missing censoring indicators we created two new observations with the same failure time and covariates, but different failure indicators and weights: the first with $\hat{\Delta}_i = 1$ and weight $\hat{p}_i$, the second with $\hat{\Delta}_i = 0$ and weight $1 - \hat{p}_i$. We used unit weight for subjects with fully observed data and recorded the estimated regression coefficient, $\hat{\beta}$. We estimated the variance of this estimate by generating 1000 bootstrap replicates of each simulated data set. We recorded the average parameter estimate, $\bar{\hat{\beta}}$ and percentile confidence intervals $(\beta_{0.025}, \beta_{0.975})$. Here, β_α is the α^{th} quantile among the 1000 bootstrap replicates.

We also compared our method to the ideal situation in which all data were observed, complete case analysis (meaning that we exclude from the dataset all observations with missing censoring indicators), and two ad-hoc methods in which we treat the missing indicators either all as censored or all as failures. Results under the assumption of MAR are included in Table 1. We estimated the bias of each method by calculating the mean difference between the estimated Cox regression coefficient and the true coefficient over the 1000 simulations. We also calculated the mean width of the confidence intervals produced by each method over the 1000 simulations. Similarly, we calculated the empirical coverage probability for the confidence intervals produced by each method by dividing the number of times that the confidence intervals contained the true value of the parameter by 1000.

[Table 1 about here.]

The empirical coverage probability using the imputed confidence interval is close to the nominal level (.95) in all simulations. Our multiple imputation method and the method of Cook and Kosorok (2004) produced unbiased estimates and valid confidence intervals in all the scenarios we considered. The estimates produced by the other methods showed a larger amount of bias and did not always approximate desired amount of coverage. Our multiple imputation method also yielded the narrowest confidence intervals in each scenario, although the method of Cook and Kosorok (2004) produced confidence intervals that were only slightly wider. Moreover, for most parameter values, the coverage probabilities for the complete case and ad hoc methods were significantly different ($p < 0.01$) from the nominal rate.

In addition, we examined the performance of our proposed methods when we changed the logistic regression model for Δ_i. We investigate two additional types of models: one in which the model contained a variable unrelated to case status and another when one variable related to case status is left out. As in the previous simulations, the failure times were generated by (1), censoring was exponential with mean 5, failure indicators were set to be missing at random, $X_{i1} \sim N(2, 1)$, $X_{i2} \sim N(\Delta_i, 0.3)$ and $Q_i = I(X_{i2} > 0.5)$ for $i = 1, \ldots, n$. We also generated $X_{i3} \sim N(0, 1)$ where X_{i1}, X_{i2}, X_{i3} were mutually independent.

In the previous simulations, we fit the data to (9) with $X_i = \{X_{i1}, X_{i2}\}$. The additional simulations instead used the covariates and parameters as follows:

(A) $\tilde{X}_i = \{1, X_{i1}, X_{i2}, X_{i3}\}$, $\tilde{\alpha} = \{\tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3\}$

(B) $\tilde{X}_i = \{1, X_{i1}\}$, $\tilde{\alpha} = \{\tilde{\alpha}_0, \tilde{\alpha}_1\}$

That is, rather than fitting model (9) to the data, we modeled the case probability with

$$logit\{Pr(\Delta_i = 1 | X_i, V_i, Q_i = 1)\} = \tilde{\alpha}' \tilde{X}_i + \gamma V_i. \tag{10}$$

Results, shown in Web Appendix A, remained similar under both alternative models when the data was missing according to mechanism II. However, the coverage was subpar when the data was missing according to mechanism I and the logistic regression model was fit according to model B. !!!FIX THIS!!! This indicates that the proposed methods are robust to alternative specification of the logistic regression model when the data is missing at random but not missing completely at random. When the data is missing completely at random, then leaving an informative covariate out of the logistic regression model decreased the performance of our method.

Finally, we conducted simulations to see how the methods compare in estimating incidence rates. Our method produced estimates much closer to the true incidence rates than those of the complete case estimate. In fact, the complete case method underestimated incidence rates by as much as a factor of 3. Details are included in Web Appendix A.

4. Data Application

In this section, we apply our method to time-to-event data in the Orofacial Pain Prospective Evaluation and Risk Assessment (OPPERA) study. OPPERA is a prospective cohort study designed to identify risk factors for first-onset TMD. A total of 3,263 initially TMD-free individuals were enrolled in OPPERA, which were recruited at four study sites from 2006-2008. TMD status was confirmed by an RDC examination (Dworkin and LeResche, 1992). For more details on the OPPERA study, see Maixner et al. (2011) and Slade et al. (2011).

Upon enrollment in the study, each OPPERA participant was assessed for a wide variety of possible risk factors for TMD, including psychological distress, previous history of painful conditions, and sensitivity to experimental pain. See Ohrbach et al. (2011), Fillingim et al. (2011), Greenspan et al. (2011), and Maixner et al. (2011) for a complete description of the baseline data sets that were collected in OPPERA. After enrollment each participant was instructed to complete a survey (called a quarterly health update or QHU) every three months. This questionnaire evaluates the frequency and severity of pain in the orofacial region during the previous three months. The QHU also evaluates pain experienced in other bodily regions, use of medications, sleep quality, psychological distress, and other possible risk factors for TMD. For a complete description of the QHU, see Slade et al. (2012).

Participants were screened as positive or negative based on whether or not their QHU responses were indicative of TMD. Participants experiencing at least 2 months with 5 or more days of orofacial pain and at least one day of pain within the past two weeks or with 5 or more days of pain in the last two weeks were considered positive on the QHU. Participants who screened positive on a QHU were asked to undergo a follow-up RDC examination by an expert dentist to determine whether or not they would be classified as an incident case of TMD.

Of the 3,263 subjects, 2,737 filled out at least 1 QHU, and the remaining 521 did not fill out any QHUs. The total number of QHUs was 28,400, of which 26,666 corresponded to subjects not yet diagnosed with TMD. (Subjects diagnosed with TMD continued to complete QHU's, but since these additional QHUs were not relevant to our time-to-event analysis, we did not analyze them.)

There were 717 positive QHUs from individuals free of TMD at the time of their QHUs. Among these positive QHU's, 486 (about 68%) received an RDC exam. In addition, upon finding a new incident case, investigators asked a matched control to come in for a clinical examination. Controls, meaning those individuals who had never been diagnosed with TMD and had negative QHUs a the time of the positive QHU leading to a diagnosis of a new case of TMD, were matched on study site, gender, and enrollment date within 2 weeks. In this manner, RDC exams were conducted on 342 matched controls. Approximately 7% of the matched controls were diagnosed with TMD. These individuals were included in our dataset as cases of TMD. This protocol resulted in 260 positive diagnoses of TMD and 556 negative diagnoses of TMD.

One of the examiners was deemed to be unreliable, so we set all of her RDC exam results to be missing and imputed them using the methods in this paper. This left 404 positive QHUs (56%) resulting in valid clinical exams. We then had 177 positive diagnoses of TMD and 530 negative diagnoses. On the individual level, after setting the exams from the unreliable examiner to missing, there were 514 people who had at least one positive QHU, 401 of which triggered positive only once, and only 39 of which had more than 2 positive QHUs.

4.1 *Hazard Ratios*

We applied our method to the OPPERA cohort to adjust for participants with missing RDC examinations. First, we estimate via logistic regression the probability of a subject being diagnosed as an incident case of TMD given a positive screening on a QHU. Due to the rich body of information collected in each QHU, we carefully selected a small number of predictor variables. Specifically, we fit a generalized linear mixed model with a logistic link function to predict the result of the RDC exam based on each item in the QHU. A mixed model was used because a sizeable (n=113) minority of subjects screened positive on multiple QHUs. All models were adjusted for study site and included a random effect term for each subject.

The majority of the variables measured on the QHU were not associated with the result of the RDC exam. The strongest predictor of a positive RDC exam was a count of facial symptoms (e.g. stiffness, fatigue, or soreness) in the previous three months. The time elapsed from baseline to the QHU was also predictive of a positive diagnosis of TMD. Several other possible predictors of a positive RDC exam were identified, but including these additional predictors in the model did not significantly improve the predictive accuracy of the model, determined by separate type III F-tests with 0.05 significance level. Thus, we estimated the probability of a positive RDC exam based on the count of facial symptoms, time since baseline, and study site. This model was used to perform multiple imputation for those who had positive QHUs but did not come in for their RDC exams. These imputed data sets were used to fit a series of Cox proportional hazards models to estimate the hazard ratio (and associated confidence interval and p-value) for each predictor using the methods described in section 2.1. Examples of predictors include perceived stress, history of comorbid chronic pain conditions and smoking status, among others.

In addition, we examined univariate relationships between the act of coming to the clinic for an RDC exam and numerous other factors. Differences between those who came in and those who did not come in were small and most were not significant. However, the fact that we found any significant associations between the baseline variables and the acts of filling out the QHUs or returning for the RDC exam indicates that the data is not missing completely at random.

Table 2 shows the results of applying our method to a subset of the putative risk factors of TMD measured in OPPERA. Due to the large number of putative risk factors measured in OPPERA, we only report a subset of these in this table. All continuous variables in the table were normalized to have mean 0 and standard deviation 1 prior to fitting the Cox models. (Thus, the hazard ratios for the continuous variables represent the hazard ratios corresponding to a one-standard deviation increase in the predictor variable.) In Table 2, all the quantitative sensory testing (QST) and psychosocial variables were continuous, and all of the clinical variables were dichotomous (and hence were not normalized).

[Table 2 about here.]

Compared to the unimputed results which treated missing censoring indicators as censored observations, imputation appeared to slightly reduce the hazard ratios for most of the psychosocial variables that were measured in OPPERA. For instance, Table 2 shows the (standardized) hazard ratios for the Pennebaker Inventory of Limbic Languidness (PILL) score, the neuroticism subscale of the Eyesenk Personality Questionnaire (EPQ), the Spielberger Trait Anxiety Inventory score, the Perceived Stress Scale, and the somatization subscale of the symptom checklist-90 (SCL90R). (See Fillingim et al. (2011) for a complete description of these psychosocial instruments.) In each case, the hazard ratios were reduced after imputation.

A similar pattern was observed after applying our imputation method to the OPPERA QST data. The mechanical pain aftersensation ratings were strongly associated with first-onset TMD before imputation, but they were only weakly associated with first-onset TMD after imputation. The pressure pain algometer ratings were also more weakly associated with TMD after imputation (and two of three ratings in Table 2 were no longer signifcantly associated with first-onset TMD at the $p < 0.05$ level). See Greenspan et al. (2011) for a more detailed description of these QST measures.

Interestingly, the hazard ratios for the presence of one or more palpation tender points at the temporalis and masseter were also attenuated after imputation. See Ohrbach et al. (2011) for a more detailed description of these variables. These tender points were evaluated as part of the RDC examination using a different protocol than the QST algometer pain ratings. However, both pain measures (algometer and palpation) were measured at the same anatomical locations on the jaw. While the palpation ratings were more strongly associated with first-onset TMD than the algometer ratings both before and after imputation, it is interesting that different pain sensitivity measures using different protocols at the same anatomical location were both attenuated by imputation.

The effects of other clinical variables were also attenuated after imputation. For example, the hazard ratios associated with being unable to open one's mouth wide in the past month and having two or more comorbid pain conditions were both noticeably attenuated after imputation. However, other clinical variables were more strongly associated with first-onset TMD after imputation. For example, having a history of at least one of five respiratory conditions was only weakly associated with first-onset TMD before imputation (HR=1.38, p=0.04), but the association was much stronger after imputation (HR=1.43, p=0.004). Also, being a current smoker was not signficantly associated with first-onset TMD before imputation (HR=1.26, p=0.24) but was associated after imputation (HR=1.49, p=0.02). See Ohrbach et al. (2011) for a more detailed description of these variables.

4.2 *Incidence Rates*

Another important aim of the OPPERA study is to calculate incidence rates of first-onset TMD. In table 3, we calculated incidence rates in two ways. First we treated all missing censoring indicators as censored. Second, we implemented the multiple imputation method in this paper. Overall rates with multiple imputation increased by nearly 2/3 compared to the unimputed rates and by 64% for females and 72% for males. Rates for whites and Hispanics were 99% and 193% higher with imputation. Thus, the incidence rate of first-onset TMD is underestimated without imputation.

[Table 3 about here.]

5. Discussion

Motivated by the research questions and framework of the OPPERA study, we presented a computationally efficient method to adjust for missing censoring indicators in time-to-event data using logistic regression and multiple imputation. Logistic regression is used to estimate the failure probability for subjects with missing censoring indicators. Then, the values are imputed and standard errors are estimated via multiple imputation. This framework is important in studies where failure status may be unknown, e.g. with interim event adjudication.

Magder and Hughes (1997) used a similar approach to estimate parameters in a logistic regression model given observed covariates when the outcome variable is measured with uncertainty by incorporating information about sensitivity and specificity. Their approach utlizes the EM algorithm to update the predicted probabilities and parameter estimates and iterating until convergence. Our assumption of MAR data renders iteration unnecessary. In fact, we conducted additional simulations, which found that our one-step procedure of estimating the predicted probabilities was nearly equivalent to the iterative procedure.

Cox proportional hazard models in the presence of missing censoring indicators were originally developed within the competing risk framework. Goetghebeur and Ryan (1995) propose competing risks proportional hazards regression models using estimating equations assuming that observations are known either to be censored or to have failed by some cause. They assume proportional hazards for each failure type and between the two failure types. For the linear transformation model, Gao and Tsiatis (2005) propose an augmented inverse probability weighted estimator and algorithm and illustrate its asymptotic and double robustness properties.

For the usual survival setting with only one cause of failure and possibly missing censoring indicators, McKeague and Subramanian (1998) and Gijbels et al. (2007) propose estimating functions for the estimation of the parameters in the Cox proportional hazards models. Subramanian (2000) conducts parameter estimation in the Cox model under the assumption of proportional censorship. Yet, these methods depend on the MCAR assumption, which may not be realistic in practice and does not hold for the OPPERA study. Chen et al. (2009) estimate Cox regression parameters using the EM algorithm and establish their consistency under basic regularity conditions, including missing at random (MAR) censoring indicators. However, their approach depends on the assumptions of piecewise constant proportional hazard functions for the censoring time as well as for the failure time.

In all of the simulations, our multiple imputation method was the most desirable method, with the narrowest valid confidence intervals and no significant bias. In particular, the method of Cook and Kosorok (2004) produced slightly wider confidence intervals in each simulation we considered. The differences were extremely small, so the performance of the two methods appear to be comparable for most practical purposes.

However, we believe our method nevertheless has several possible advantages over the method of Cook and Kosorok (2004). First, bootstrapping is much more intensive computationally than our multiple imputation approach. Calculating bootstrap confidence intervals generally requires at least 1000 bootstrap replicates (Efron and Tibshirani, 1993), whereas as few as 10 imputed data sets may be sufficient for multiple imputation (Little and Rubin, 2002). Although the difference in the computing time of the two methods is small for a single fitted model, many such models will need to be fitted in the course of the OPPERA study. OPPERA has already collected data on approximately 3000 genetic markers and has plans to collect data on approximately a million genetic markers in a genome-wide association study. Thus, at least 3000 (and potentially as many as a million) Cox models will need to be fit, and our proposed method may allow for a significant decrease in computing time. Moreover, our method can also be easily implemented in popular statistical software packages (such as SAS) without additional programming.

Additionally, our methodology may easily be extended to other survival models, such as Poisson regression. We conducted simulations (Table S9 in the Web Appendix) that produced similar results for Poisson regression compared to the Cox model. Our extension to Poisson regression allows for estimation of incidence rates, which comprise one of the research aims of the OPPERA study. In particular, estimates for failure rates were biased when missing censoring indicators were treated as censored or when the complete case method was used and unbiased when we employed the methodology in this paper.

The OPPERA protocol also included a baseline case-control study, which allowed us to estimate the false negative rate of the QHU. Whenever a new incident case of TMD was diagnosed, a subject from the same study site who had screened negative on a QHU in the same quarter was examined for TMD. About 5% of these matched controls had TMD, i.e. there was a false negative rate of 5%. In order to parallel the set-up of OPPERA, in all simulations, we set $\Delta_i = 0$ if $Q_i = 0$, indicating a negative QHU. This is because, assuming subject i was not a matched control, he or she would not be subjected to an RDC exam, and therefore would not have been considered a case of TMD.

The assumption of missing at random seems reasonable for the OPPERA study. However, our methodology can be generalized for non-ignorable missing censoring indicators by using the EM-algorithm to fit a logistic regression model, estimate predicted probabilities, refit the data to a model weighted by the predicted probabilities and iterating until convergence. This is a topic of future study.

In the OPPERA dataset, hazard ratios, confidence intervals, and p-values differed noticeably with and without imputation for a number of variables. Although some of the results remained qualitatively unchanged based on whether or not our methodology was utilized, we note that even small changes in hazard ratios are important. In addition, estimated incidence rates were significantly increased after imputation. Since the results of OPPERA may become normative in the orofacial pain literature, precise calculation of the incidence rate of TMD and the hazard ratios associated with putative risk factors is important. Thus, imputation is recommended.

Acknowledgements

The authors would like to acknowledge and thank the principal investigators of the OPPERA study, namely William Maixner, Luda Diatchenko, Bruce Weir, Richard Ohrbach, Roger Fillingim, Joel Greenspan, and Ronald Dubner. The OPPERA study was supported by NIH/NIDCR grant U01DE017018. The first author was also supported by NIH/NIEHS T32ES007018.

Supplementary Materials

Web Appendix A and Web Tables 1-13, referenced in Section 3, are available with this paper at the Biometrics website on Wiley Online Library

References

Bair, E., Brownstein, N. C., Ohrbach, R., Greenspan, J. D., Dubner, R., Fillingim, R. B., Maixner, W., Smith, S., Diatchenko, L., Gonzalez, Y., Gordon, S., Lim, P.-F., Ribeiro-Dasilva, M., Dampier, D., Knott, C., and Slade, G. D. (2013b). Study design, methods, sample characteristics and loss-to-follow-up: the oppera prospective cohort study. Submitted.

Bair, E., Brownstein, N. C., Ohrbach, R., Greenspan, J. D., Dubner, R., Fillingim, R. B., Maixner, W., Smith, S., Diatchenko, L., Gonzalez, Y., Gordon, S., Lim, P.-F., Ribeiro-Dasilva, M., Dampier, D., Knott, C., and Slade, G. D. (2013a). Supplementary materials to: "study design, methods, sample characteristics and loss-to-follow-up: the oppera prospective cohort study". Submitted.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine* **24,** 1713–1723.

Chen, P., He, R., Shen, J.-s., and Sun, J.-g. (2009). Regression analysis of right-censored failure time data with missing censoring indicators. *Acta Mathematicae Applicatae Sinica (English Series)* **25,** 415–426. 10.1007/s10255-008-8807-1.

Cook, T. D. and Kosorok, M. R. (2004). Analysis of time-to-event data with incomplete event adjudication. *Journal of the American Statistical Association* **99,** pp. 1140–1152.

Dworkin, S. and LeResche, L. (1992). Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. *Journal of Craniomandibular Disorders* **6,** 301–355.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Chapman and Hall/CRC, Boca Raton, FL.

Fillingim, R. B., Ohrbach, R., Greenspan, J. D., Knott, C., Dubner, R., Bair, E., Baraian, C., Slade, G. D., and Maixner, W. (2011). Potential psychosocial risk factors for chronic tmd: Descriptive data and empirically identified domains from the oppera case-control study. *The Journal of Pain* **12,** T46 – T60.

Gao, G. and Tsiatis, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika* **92,** pp. 875–891.

Gijbels, I., Lin, D., and Ying, Z. (2007). Non- and semi- parametric analysis of failure-time data with missing failure indicators. *Lecture Notes-Monograph Series* **54,** 203–223.

Goetghebeur, E. and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika* **82,** 821–833.

Greenspan, J. D., Slade, G. D., Bair, E., Dubner, R., Fillingim, R. B., Ohrbach, R., Knott, C., Mulkey, F., Rothwell, R., and Maixner, W. (2011). Pain sensitivity risk factors for chronic tmd: Descriptive data and empirically identified domains from the oppera case control study. *The Journal of Pain* **12,** T61 – T74.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Wiley Series in Probability and Mathematical Statistics.

Magder, L. S. and Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* **146,** 195–203.

Maixner, W., Diatchenko, L., Dubner, R., Fillingim, R. B., Greenspan, J. D., Knott, C., Ohrbach, R., Weir, B., and Slade, G. D. (2011). Orofacial pain prospective evaluation and risk assessment study the oppera study. *The Journal of Pain* **12,** T4–T11.

Maixner, W., Greenspan, J. D., Dubner, R., Bair, E., Mulkey, F., Miller, V., Knott, C., Slade,

G. D., Ohrbach, R., Diatchenko, L., and Fillingim, R. B. (2011). Potential autonomic risk factors for chronic tmd: Descriptive data and empirically identified domains from the oppera case-control study. *The Journal of Pain* **12,** T75 – T91.

McKeague, I. W. and Subramanian, S. (1998). Product-limit estimators and cox regression with missing censoring information. *Scandinavian Journal of Statistics* **25,** pp. 589–601.

Ohrbach, R., Fillingim, R. B., Mulkey, F., Gonzalez, Y., Gordon, S., Gremillion, H., Lim, P.-F., Ribeiro-Dasilva, M., Greenspan, J. D., Knott, C., Maixner, W., and Slade, G. (2011). Clinical findings and pain symptoms as potential risk factors for chronic tmd: Descriptive data and empirically identified domains from the oppera case-control study. *The Journal of Pain* **12,** T27 – T45.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** pp. 581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91,** pp. 473–489.

Slade, G. D., Bair, E., By, K., Mulkey, F., Baraian, C., Rothwell, R., Reynolds, M., Miller, V., Gonzalez, Y., Gordon, S., Ribeiro-Dasilva, M., Lim, P. F., Greenspan, J. D., Dubner, R., Fillingim, R. B., Diatchenko, L., Maixner, W., Dampier, D., Knott, C., and Ohrbach, R. (2011). Study methods, recruitment, sociodemographic findings, and demographic representativeness in the oppera study. *The Journal of Pain* **12,** T12–T26.

Slade, G. D., Sanders, A., Bair, E., Brownstein, N. C., Fillingim, R. B., Maixner, W., Greenspan, J. D., and Ohrbach, R. (2012). Pre-clinical episodes of orofacial pain symptoms and their association with healthcare behaviors in the oppera prospective cohort study. Accepted.

Subramanian, S. (2000). Efficient estimation of regression coefficients and baseline hazard under proportionality of conditional hazards. *Journal of Statistical Planning and Inference* **84,** 81 – 94.

Table 1
Simulation Results for MAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0005	0.0006	0.1668	0.0001	0.938
	Complete Case	0.0018	0.0007	0.2155	0.0001	0.951
	Treat all as Censored	0.1053	0.0007	0.2131	0.0001	0.494
	Treat all as Failures	0.0018	0.0006	0.1701	0.0001	0.941
	Cook & Kosorok	-0.0010	0.0006	0.1738	0.0001	0.943
	Multiple Imputation	-0.0010	0.0006	0.1716	0.0001	0.938
-1.5	Full Data	-0.0008	0.0010	0.3185	0.0002	0.966
	Complete Case	-0.0626	0.0015	0.4343	0.0003	0.930
	Treat all as Censored	0.1215	0.0014	0.4229	0.0003	0.778
	Treat all as Failures	0.0680	0.0010	0.3160	0.0002	0.852
	Cook & Kosorok	0.0002	0.0011	0.3412	0.0004	0.951
	Multiple Imputation	0.0001	0.0011	0.3309	0.0002	0.961
-3	Full Data	-0.0301	0.0025	0.7627	0.0009	0.957
	Complete Case	-0.1996	0.0037	1.0840	0.0017	0.913
	Treat all as Censored	0.0987	0.0035	1.0417	0.0016	0.919
	Treat all as Failures	0.5875	0.0024	0.6307	0.0006	0.104
	Cook & Kosorok	-0.0275	0.0027	0.9112	0.0017	0.946
	Multiple Imputation	-0.0282	0.0027	0.8059	0.0011	0.947

*: The Monte Carlo error is 0.007.

Table 2
Results from the OPPERA Study

Consider All MCIs as Censored					**Multiple Imputation**			
	HR	LCL	UCL	P	HR	LCL	UCL	P
Clinical Variable								
In the last month could not open mouth wide	3.26	1.83	5.84	<0.0001	2.45	1.42	4.22	0.0012
Has two or more comorbid chronic pain disorders	3.08	2.26	4.21	<0.0001	2.50	1.90	3.29	<0.0001
History of 5 respiratory conditions	1.38	1.01	1.87	0.0408	1.45	1.13	1.87	0.0040
Smoking: current	1.26	0.86	1.84	0.2403	1.49	1.07	2.09	0.0199
Smoking: former	1.87	1.22	2.87	0.0041	1.65	1.12	2.43	0.0106
One or more palpation tender points: right temporalis	1.83	1.32	2.52	0.0002	1.54	1.18	2.02	0.0017
One or more palpation tender points: left temporalis	1.60	1.14	2.25	0.0064	1.48	1.12	1.97	0.0060
One or more palpation tender points: right masseter	1.85	1.35	2.53	0.0001	1.63	1.25	2.12	0.0003
One or more palpation tender points: left masseter	1.70	1.23	2.35	0.0013	1.53	1.17	2.01	0.0021
Quantitative Sensory Testing Variable								
Pressure pain threshold: temporalis	1.26	1.07	1.49	0.0065	1.16	1.01	1.33	0.0335
Pressure pain threshold: masseter	1.23	1.04	1.45	0.0170	1.15	1.00	1.32	0.0576
Pressure pain threshold: TM joint	1.25	1.05	1.48	0.0106	1.14	1.00	1.30	0.0555
Mechanical pain aftersensation: 512mN probe, 15 s	1.23	1.09	1.38	0.0006	1.15	1.03	1.28	0.0123
Mechanical pain aftersensation: 512mN probe, 30 s	1.20	1.07	1.34	0.0020	1.12	1.01	1.25	0.0328
Psychosocial Variable								
PILL Global Score	1.52	1.35	1.71	<0.0001	1.46	1.31	1.62	<0.0001
EPQ-R Neuroticism	1.39	1.21	1.60	<0.0001	1.26	1.12	1.42	0.0002
Trait Anxiety Inventory	1.43	1.25	1.64	<0.0001	1.35	1.21	1.52	<0.0001
Perceived Stress Scale	1.35	1.17	1.55	<0.0001	1.30	1.16	1.47	<0.0001
SCL 90R Somatization	1.44	1.31	1.58	<0.0001	1.40	1.29	1.52	<0.0001

Table 3
Estimated TMD Incidence Rates (in Percentages) With and Without Imputation

	No MI	MI	Percent Change
Overall	2.23	3.70	66
Males	1.87	3.22	72
Females	2.46	4.03	64
White	1.70	3.37	99
Black	4.20	5.32	27
Hispanic	1.17	3.44	193
Other	1.10	1.80	63

S1. Overview of Additional Simulations

In this appendix, we provide the results of additional simulations. We investigate the performance of the method under a variety of missing data mechanisms as well as when we specify alternative logistic regression models for the probability of being a case given that the participant screened positive on the simple examination.

Recall that we created missing censoring indicators under the following classical missing data mechanisms of Rubin (1976):

(I) The probability of having a missing censoring indicator is independent of the data. This is known as missing completely at random (MCAR).

(II) The probability of having a missing censoring indicator depends on an observed covariate. This is known as missing at random (MAR).

(III) The probability of having a missing censoring indicator depends on the censoring indicator. This is known as missing not at random (MNAR).

For mechanism I, we randomly set 40% of the censoring indicators, $\Delta_i = I(T_i < C_i)$, to be missing. For mechanism III, we investigated two secarios.

(A) In the first, we set 30% of the censored observations and 50% of the failures to have missing indicators.

(B) In the second, we set 20% of the censored observations and 60% of the failures to have missing indicators.

S2. Simulations Under MCAR

[Table S1 about here]

When the data were MCAR, our method had less bias on average than the complete case method depended on the true parameter value. Not only did our method have adequate coverage, but it had the most narrow confidence intervals of the methods with adequate coverage. As in other simulations, the method treating all missing indicators as failures had poor coverage and introduced extreme bias. The complete case method and the method that treat all missing censoring indicators as censored were valid, but had much wider confidence intervals than our method.

However, note that the complete case method would not be applicable to the OPPERA study. According to the OPPERA protocol, participants who do not screen positive on the QHU are automatically considered censored. Only participants who screen positive on the QHU (i.e. those with $Q = 1$) may potentially have missing censoring indicators.

S3. Additional Simulations Under MAR

In order to more closely parallel the OPPERA study, we simulated data for which we randomly set to missing 40% of the censoring indicators for those with $Q = 1$. This setup corresponds to the still strict assumption that the probability that a participant completes an RDC exam depends only on whether or not their QHU was positive. The logistic regression model in this case included the covariates (X_{i1} and X_{i2}) as before, but not the time of the QHU. Results are included in Table S2. All methods had a negligible amount of bias in these scenarios except for the complete case method and the method that treated all missing indicators as failures. In these simulations, the complete case method also displayed extreme bias and poor coverage. This indicates that a complete case analysis would not be appropriate for a study such as OPPERA.

[Table S2 about here]

S4. Alternative Logistic Regression Models

We investigated alternate logistic regression models for the probability of being a case given covariates. Recall that we originally modeled the probability of being a case as

$$p(\Delta_i = 1 | X_i, \alpha) = \frac{\exp(\alpha' X_i)}{1 + \exp(\alpha' X_i)} \tag{11}$$

The alternative models were of the form (11) but used the covariates

(1) $\tilde{X}_i = \{X_{i1}, X_{i2}, X_{i3}\}$ and V_i

(2) $\tilde{X}_i = \{X_{i1}\}$ and V_i

The original logistic model had the covariates $X_i = \{X_{i1}, X_{i2}\}$ and V_i where $X_{i1} \sim N(0, 2)$, $X_{i1} \sim N(0, \Delta_i)$, $X_{i3} \sim N(0, 1)$ are mutually independent for $j = 1, 2, 3$ and $i = 1, \ldots, n$.

[Table S3 about here]

[Table S4 about here]

[Table S5 about here]

[Table S6 about here]

S5. Simulations Under MNAR

[Table S7 about here]

[Table S8 about here]

When the data were MNAR, bias increased for all methods. In particular, the complete case method consistently displayed a high amount of bias and did not achieve the desired coverage rate. For our imputation method and the method of Cook and Kosorok, bias increased and coverage decreased as the true parameter value increased. This indicates that when the MAR assumption is violated, our method as well as the method of Cook and Kosorok (2004) may not be valid. On the other hand, even when the data was not missing at random, our method provided an improvement in terms of bias and coverage over the complete case method and the method that treats all missing subjects as failures. Moreover, coverage was slightly greater for our method than for the method of Cook and Kosorok (2004).

S6. Simulations for Poisson Regression

We investigated the performance of our method if the desired time-to-event analysis was a Poisson regression model rather than a Cox model. Poisson models are commonly used to estimate incidence rates, which were desired in the OPPERA study. The following section details the simulations.

The simulations were identical to those in the main paper, with the exception of fitting the imputed data to Poisson regression models rather than to Cox proportional hazards models. That is, we fit the data from imputation $j = 1, \ldots, m$ to the model

$$log(\mu_i) = \alpha + \beta x_{i1} + log(V_i). \tag{12}$$

where μ_i is the expected number of cases and the offset, $log(V_i)$, is the logarithm of the survival time. We measured the bias, defined as $\hat{\beta}$ minus the true value, where $\beta \in \{-0.5, -1.5, -3\}$.

The Cook and Kosorok method does not immediately generalize to Poisson regression.

Consequently, we only compared our method to the unachievable ideal of no missing data, the complete case method, and the two ad-hoc methods.

The use of Poisson regression allows us to estimate incidence rates. For each simulation, we estimate the incidence rate based on the fit of the Poisson regression model in (12). Specifically, estimated incidence rates for fixed values of X_{i1} are given by

$$exp(\alpha + \beta x_{i1}) \tag{13}$$

We present incidence rates for the quartiles of the random variable X_{i1}, i.e. the quartiles of the $N(2, 1)$ distribution.

Just as when Cox regression was used, our method had coverage close to the nominal rate when Poisson regression was used. None of the other methods had proper coverage for all of the simulations. Multiple imputation yielded the least bias of all the methods besides the unacheivable ideal of observing all data and produced more narrow confidence intervals than the complete case method and the method that treats all missing censoring indicators as censored.

The bias evident in parameter estimation was compounded for incidence rates. Complete case and analyses consistently underestimated incidence. In fact, the complete case method underestimated incidence by about 30-200%. By contrast, our method differed from the unachievable ideal by only about 4-6%. See Table S10. In conclusion, our method is the only valid method to estimate parameters and incidence rates from Poisson regression.

[Table S9 about here]

[Table S10 about here]

Received March 2013. *Revised June* 2013. *Accepted September* 2013.

Table S1
Simulation Results for MCAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0010	0.0006	0.1671	0.0001	0.930
	Complete Case	-0.0042	0.0007	0.2218	0.0001	0.956
	Treat all as Censored	-0.0021	0.0007	0.2212	0.0001	0.947
	Treat all as Failures	0.0932	0.0005	0.1525	0.0001	0.323
	Cook & Kosorok	-0.0005	0.0006	0.1784	0.0002	0.943
	Multiple Imputation	-0.0005	0.0006	0.1714	0.0001	0.933
-1.5	Full Data	-0.0008	0.0010	0.3185	0.0002	0.966
	Complete Case	-0.0056	0.0014	0.4261	0.0003	0.948
	Treat all as Censored	-0.0025	0.0014	0.4229	0.0003	0.950
	Treat all as Failures	0.8277	0.0007	0.2036	0.0001	0.000
	Cook & Kosorok	-0.0003	0.0011	0.3512	0.0004	0.952
	Multiple Imputation	-0.0004	0.0011	0.3306	0.0002	0.949
-3	Full Data	-0.0190	0.0025	0.7574	0.0008	0.952
	Complete Case	-0.0321	0.0036	1.0255	0.0016	0.942
	Treat all as Censored	-0.0205	0.0035	1.0070	0.0015	0.950
	Treat all as Failures	2.5225	0.0009	0.2216	0.0001	0.000
	Cook & Kosorok	-0.0229	0.0028	0.9459	0.0024	0.953
	Multiple Imputation	-0.0239	0.0028	0.7961	0.0010	0.933

*: The Monte Carlo error is 0.007.

Table S2
Additional Simulation Results for MAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0010	0.0006	0.1671	0.0001	0.930
	Complete Case	-0.0543	0.0008	0.2210	0.0001	0.825
	Treat all as Censored	-0.0021	0.0007	0.2212	0.0001	0.947
	Treat all as Failures	0.0041	0.0006	0.1703	0.0001	0.929
	Cook & Kosorok	-0.0012	0.0006	0.1742	0.0001	0.934
	Multiple Imputation	-0.0012	0.0006	0.1722	0.0001	0.938
-1.5	Full Data	-0.0008	0.0010	0.3185	0.0002	0.966
	Complete Case	-0.1329	0.0014	0.4283	0.0003	0.759
	Treat all as Censored	-0.0025	0.0014	0.4229	0.0003	0.950
	Treat all as Failures	0.0849	0.0011	0.3147	0.0002	0.790
	Cook & Kosorok	-0.0006	0.0011	0.3411	0.0004	0.951
	Multiple Imputation	-0.0006	0.0011	0.3320	0.0002	0.957
-3	Full Data	-0.0190	0.0025	0.7574	0.0008	0.952
	Complete Case	-0.2342	0.0037	1.0336	0.0016	0.883
	Treat all as Censored	-0.0205	0.0035	1.0070	0.0015	0.950
	Treat all as Failures	0.6626	0.0025	0.6165	0.0006	0.047
	Cook & Kosorok	-0.0232	0.0028	0.8775	0.0016	0.940
	Multiple Imputation	-0.0240	0.0028	0.7996	0.0010	0.937

*: The Monte Carlo error is 0.007.

Table S3
Alternative model of type (1), MCAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	0.0009	0.0005	0.1668	0.0001	0.960
	Complete Case	0.0006	0.0007	0.2216	0.0001	0.960
	Treat all as Censored	0.0023	0.0007	0.2208	0.0001	0.964
	Treat all as Failures	0.0946	0.0005	0.1522	0.0001	0.334
	Cook & Kosorok	0.0020	0.0006	0.1807	0.0002	0.963
	Multiple Imputation	0.0021	0.0006	0.1708	0.0001	0.955
-1.5	Full Data	-0.0050	0.0011	0.3187	0.0002	0.943
	Complete Case	-0.0077	0.0014	0.4245	0.0004	0.949
	Treat all as Censored	-0.0038	0.0014	0.4218	0.0003	0.947
	Treat all as Failures	0.8225	0.0007	0.2040	0.0001	0.000
	Cook & Kosorok	-0.0041	0.0011	0.3526	0.0004	0.942
	Multiple Imputation	-0.0041	0.0011	0.3306	0.0002	0.940
-3	Full Data	-0.0191	0.0026	0.7586	0.0008	0.950
	Complete Case	-0.0389	0.0035	1.0301	0.0016	0.949
	Treat all as Censored	-0.0276	0.0034	1.0108	0.0014	0.947
	Treat all as Failures	2.5217	0.0008	0.2214	0.0001	0.000
	Cook & Kosorok	-0.0155	0.0029	0.9686	0.0025	0.963
	Multiple Imputation	-0.0157	0.0028	0.7977	0.0011	0.940

*: The Monte Carlo error is 0.007.

Table S4
Alternative model of type (1), MAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	0.0009	0.0005	0.1668	0.0001	0.960
	Complete Case	0.0060	0.0007	0.2155	0.0001	0.952
	Treat all as Censored	0.1084	0.0007	0.2130	0.0001	0.480
	Treat all as Failures	0.0036	0.0005	0.1702	0.0001	0.956
	Cook & Kosorok	0.0009	0.0005	0.1739	0.0001	0.960
	Multiple Imputation	0.0009	0.0005	0.1717	0.0001	0.959
-1.5	Full Data	-0.0050	0.0011	0.3187	0.0002	0.943
	Complete Case	-0.0663	0.0014	0.4332	0.0004	0.914
	Treat all as Censored	0.1178	0.0014	0.4219	0.0003	0.797
	Treat all as Failures	0.0645	0.0011	0.3163	0.0002	0.861
	Cook & Kosorok	-0.0040	0.0011	0.3415	0.0004	0.941
	Multiple Imputation	-0.0040	0.0011	0.3311	0.0002	0.946
-3	Full Data	-0.0191	0.0026	0.7586	0.0008	0.950
	Complete Case	-0.2009	0.0038	1.0867	0.0017	0.918
	Treat all as Censored	0.1055	0.0035	1.0403	0.0015	0.911
	Treat all as Failures	0.6054	0.0025	0.6266	0.0006	0.092
	Cook & Kosorok	-0.0170	0.0029	0.9122	0.0017	0.949
	Multiple Imputation	-0.0158	0.0029	0.8066	0.0012	0.934

*: The Monte Carlo error is 0.007.

Table S5
Alternative model of type (2), MCAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0001	0.0006	0.1667	0.0001	0.941
	Complete Case	0.0011	0.0008	0.2211	0.0001	0.939
	Treat all as Censored	0.0009	0.0007	0.2205	0.0001	0.939
	Treat all as Failures	0.0938	0.0005	0.1522	0.0001	0.331
	Cook & Kosorok	0.0840	0.0005	0.1581	0.0001	0.484
	Multiple Imputation	0.0840	0.0005	0.1551	0.0001	0.440
-1.5	Full Data	-0.0046	0.0011	0.3187	0.0002	0.941
	Complete Case	-0.0068	0.0014	0.4242	0.0004	0.939
	Treat all as Censored	-0.0029	0.0014	0.4217	0.0003	0.945
	Treat all as Failures	0.8236	0.0007	0.2041	0.0001	0.000
	Cook & Kosorok	0.5062	0.0013	0.3936	0.0006	0.003
	Multiple Imputation	0.5056	0.0013	0.2584	0.0002	0.000
-3	Full Data	-0.0229	0.0026	0.7606	0.0009	0.954
	Complete Case	-0.0335	0.0034	1.0296	0.0016	0.965
	Treat all as Censored	-0.0282	0.0034	1.0132	0.0014	0.956
	Treat all as Failures	2.5266	0.0008	0.2212	0.0001	0.000
	Cook & Kosorok	0.8890	0.0034	1.0710	0.0026	0.191
	Multiple Imputation	0.8856	0.0034	0.5425	0.0008	0.026

*: The Monte Carlo error is 0.007.

Table S6
Alternative model of type (2), MAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0034	0.0005	0.1671	0.0001	0.950
	Complete Case	-0.0012	0.0007	0.2157	0.0001	0.943
	Treat all as Censored	0.1005	0.0007	0.2133	0.0001	0.544
	Treat all as Failures	-0.0008	0.0006	0.1704	0.0001	0.955
	Cook & Kosorok	-0.0036	0.0006	0.1739	0.0001	0.951
	Multiple Imputation	-0.0036	0.0006	0.1721	0.0001	0.953
-1.5	Full Data	-0.0055	0.0010	0.3183	0.0002	0.955
	Complete Case	-0.0638	0.0014	0.4322	0.0004	0.925
	Treat all as Censored	0.1177	0.0014	0.4218	0.0003	0.787
	Treat all as Failures	0.0625	0.0010	0.3161	0.0002	0.869
	Cook & Kosorok	-0.0054	0.0011	0.3451	0.0004	0.944
	Multiple Imputation	-0.0055	0.0011	0.3359	0.0002	0.952
-3	Full Data	-0.0229	0.0026	0.7606	0.0009	0.954
	Complete Case	-0.1961	0.0037	1.0861	0.0017	0.931
	Treat all as Censored	0.1006	0.0035	1.0436	0.0015	0.902
	Treat all as Failures	0.5975	0.0025	0.6285	0.0006	0.093
	Cook & Kosorok	-0.0222	0.0029	0.9052	0.0017	0.929
	Multiple Imputation	-0.0244	0.0029	0.8246	0.0011	0.940

*: The Monte Carlo error is 0.007.

Table S7
Simulation Results for MNAR, scenario A

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0021	0.0006	0.1671	0.0001	0.938
	Complete Case	-0.0755	0.0008	0.2424	0.0004	0.778
	Treat all as Censored	-0.0022	0.0008	0.2421	0.0001	0.942
	Treat all as Failures	0.0023	0.0006	0.1705	0.0001	0.942
	Cook & Kosorok	-0.0054	0.0006	0.1760	0.0001	0.940
	Multiple Imputation	-0.0054	0.0006	0.1732	0.0001	0.943
-1.5	Full Data	-0.0030	0.0011	0.3185	0.0002	0.942
	Complete Case	-0.1744	0.0016	0.4691	0.0004	0.717
	Treat all as Censored	-0.0049	0.0015	0.4623	0.0004	0.947
	Treat all as Failures	0.0646	0.0011	0.3172	0.0002	0.875
	Cook & Kosorok	-0.0206	0.0011	0.346	0.0004	0.921
	Multiple Imputation	-0.0207	0.0011	0.3362	0.0002	0.939
-3	Full Data	-0.0289	0.0026	0.7611	0.0009	0.948
	Complete Case	-0.3308	0.0040	1.1490	0.0018	0.824
	Treat all as Censored	-0.0339	0.0039	1.1060	0.0016	0.935
	Treat all as Failures	0.5242	0.0026	0.6487	0.0007	0.181
	Cook & Kosorok	-0.0737	0.0029	0.9026	0.0018	0.898
	Multiple Imputation	-0.0745	0.0029	0.8194	0.0011	0.940

*: The Monte Carlo error is 0.007.

Table S8
Simulation Results for MNAR, scenario B

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0021	0.0006	0.1671	0.0001	0.938
	Complete Case	-0.0998	0.0009	0.2713	0.0002	0.687
	Treat all as Censored	-0.0016	0.0009	0.2707	0.0001	0.943
	Treat all as Failures	0.0009	0.0006	0.1707	0.0001	0.943
	Cook & Kosorok	-0.0094	0.0006	0.1783	0.0001	0.933
	Multiple Imputation	-0.0095	0.0006	0.1747	0.0001	0.934
-1.5	Full Data	-0.0030	0.0011	0.3185	0.0002	0.942
	Complete Case	-0.2278	0.0018	0.5289	0.0005	0.618
	Treat all as Censored	-0.0046	0.0017	0.5169	0.0004	0.958
	Treat all as Failures	0.0444	0.0011	0.3201	0.0002	0.910
	Cook & Kosorok	-0.0398	0.0012	0.3534	0.0004	0.902
	Multiple Imputation	-0.0398	0.0012	0.3417	0.0002	0.920
-3	Full Data	-0.0289	0.0026	0.7611	0.0009	0.948
	Complete Case	-0.4316	0.0046	1.3085	0.0023	0.781
	Treat all as Censored	-0.0464	0.0044	1.2434	0.002	0.932
	Treat all as Failures	0.3661	0.0027	0.6850	0.0007	0.451
	Cook & Kosorok	-0.1180	0.0030	0.9425	0.0021	0.870
	Multiple Imputation	-0.1189	0.0030	0.8394	0.0012	0.924

*: The Monte Carlo error is 0.007.

Table S9
Simulation Results for Poisson Models, MAR

β	Method	Bias	SE (Bias)	Width	SE (Width)	Coverage*
-0.5	Full Data	-0.0036	0.0005	0.1596	0.0001	0.956
	Complete Case	-0.0101	0.0007	0.2067	0.0001	0.948
	Treat all as Censored	0.0813	0.0007	0.2049	0.0001	0.652
	Treat all as Failures	-0.0002	0.0006	0.1628	0.0001	0.957
	Multiple Imputation	-0.0036	0.0006	0.1642	0.0001	0.959
-1.5	Full Data	-0.0005	0.0010	0.2864	0.0002	0.945
	Complete Case	-0.1008	0.0015	0.3956	0.0004	0.829
	Treat all as Censored	0.0704	0.0014	0.3883	0.0003	0.898
	Treat all as Failures	0.0717	0.001	0.2857	0.0002	0.820
	Multiple Imputation	0.0006	0.0011	0.2979	0.0002	0.940
-3	Full Data	-0.0184	0.0021	0.5994	0.0007	0.958
	Complete Case	-0.2733	0.0032	0.871	0.0016	0.792
	Treat all as Censored	0.0206	0.003	0.8485	0.0012	0.954
	Treat all as Failures	0.5232	0.0025	0.5544	0.0006	0.085
	Multiple Imputation	-0.0219	0.0024	0.6353	0.0009	0.940

*: The Monte Carlo error is 0.007.

Table S10
Simulation Results for Incidence Rates

β	Method	Q1	SE	Q2	SE	Q3	SE
-0.5	Full Data	0.5157	0.0003	0.3671	0.0002	0.2615	0.0002
	Complete Case	0.4261	0.0004	0.3019	0.0002	0.2142	0.0002
	Treat all as Censored	0.2991	0.0003	0.2254	0.0002	0.1701	0.0002
	Treat all as Failures	0.4949	0.0003	0.3531	0.0002	0.2521	0.0002
	Multiple Imputation	0.4910	0.0003	0.3495	0.0002	0.2490	0.0002
-1.5	Full Data	0.1375	0.0001	0.0501	0.0001	0.0183	0.0000
	Complete Case	0.0968	0.0001	0.0330	0.0001	0.0113	0.0000
	Treat all as Censored	0.0753	0.0001	0.0288	0.0000	0.0110	0.0000
	Treat all as Failures	0.1388	0.0001	0.0530	0.0001	0.0203	0.0000
	Multiple Imputation	0.1309	0.0001	0.0477	0.0001	0.0174	0.0000
-3	Full Data	0.0186	0.0000	0.0025	0.0000	0.0003	0.0000
	Complete Case	0.0106	0.0000	0.0012	0.0000	0.0001	0.0000
	Treat all as Censored	0.0097	0.0000	0.0013	0.0000	0.0002	0.0000
	Treat all as Failures	0.0301	0.0001	0.0058	0.0000	0.0011	0.0000
	Multiple Imputation	0.0176	0.0000	0.0023	0.0000	0.0003	0.0000

*: Q1 denotes rates based on the lower quartile of X_{i1}.

*: Q2 denotes rates based on the median of X_{i1}.

*: Q3 denotes rates based on the upper quartile of X_{i1}.